Metabarcoding and Metagenomics 10: 235–260 (2026)
DOI: 10.3897/mbmg.10.173612

Forum Paper

Navigating the European aquatic eDNA landscape: Opportunities for metadata standardisation and data mobilisation

Ioannis Kavakiotis[1*], Peter Woollard[2*], Camila Babo[3,4], Elianne Egge[5], Dawid Krawczyk[6], Tiina Laamanen[7], Joana Paupério[2], Karolina Bacela-Spychalska[6], Nicolas Pade[8], Filipe O. Costa[9,10], Michal Grabowski[6], Veera Norros[11], Kristian Meissner[12], Spiros Papakostas[1]

1 *Department of Science and Technology, School of Science and Technology, International Hellenic University, 57001 Thessaloniki, Greece*
2 *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, Cambridge, UK*
3 *CIBIO-InBIO, Research Center in Biodiversity and Genetic Resources, 4485-661 Vairão, Portugal*
4 *BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal*
5 *Norwegian Institute for Water Research, Økernveien 94, N-0579 Oslo, Norway*
6 *Department of Invertebrate Zoology and Hydrobiology, University of Lodz, Banacha 12/16, 90-237 Łódź, Poland*
7 *Finnish Environment Institute, Syke, Paavo Havaksentie 3, 90540 Oulu, Finland*
8 *European Marine Biological Resource Centre (EMBRC-ERIC), Paris, France*
9 *Centre of Molecular and Environmental Biology (CBMA) and ARNET-Aquatic Research Network, University of Minho, Campus de Gualtar, Braga, Portugal*
10 *Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, Campus de Gualtar, Braga, Portugal*
11 *Finnish Environment Institute (Syke), Latokartanonkaari 11, 00790 Helsinki, Finland*
12 *Finnish Environment Institute (Syke), Survontie 9a, 40100 Jyväskylä, Finland*
Corresponding author: Spiros Papakostas (spapakostas@ihu.edu.gr)

OPEN ACCESS

Academic editor: Miwa Takahashi
Received: 30 September 2025
Accepted: 26 February 2026
Published: 20 March 2026

Citation: Kavakiotis I, Woollard P, Babo C, Egge E, Krawczyk D, Laamanen T, Paupério J, Bacela-Spychalska K, Pade N, Costa FO, Grabowski M, Norros V, Meissner K, Papakostas S (2026) Navigating the European aquatic eDNA landscape: Opportunities for metadata standardisation and data mobilisation. Metabarcoding and Metagenomics 10: e173612. https://doi.org/10.3897/mbmg.10.173612

Copyright: © Ioannis Kavakiotis et al.
This is an open access article distributed under terms of the Creative Commons Attribution License (Attribution 4.0 International – CC BY 4.0).

Abstract

Environmental DNA (eDNA) has emerged as a transformative tool for monitoring aquatic biodiversity, offering a non-invasive and highly sensitive approach to detecting organisms across diverse ecosystems. However, its effective downstream application across Europe in environmental management is hindered by inconsistencies in data standardisation, metadata reporting, and accessibility. This perspective comprehensively evaluates current data repositories, data submission workflows, and standardisation efforts within the European aquatic eDNA landscape. By employing a multi-method approach, including an inventory of eDNA databases, a metadata assessment, a stakeholder questionnaire, and a generative Artificial Intelligence (AI)-driven analysis of scientific literature, our findings reveal substantial variability in metadata reporting practices, with several areas misaligned with Findable, Accessible, Interoperable, and Reusable (FAIR) principles. While some repositories demonstrate strong data curation and accessibility, others lack essential metadata descriptors, limiting interoperability. We identify critical gaps in metadata submission, particularly concerning sampling methods and wet lab workflows, which heavily impact data reusability. The use of generative AI in this study further enabled large-scale identification of recurring reporting weaknesses, highlighting structural challenges that extend beyond individual studies. Addressing these gaps and leveraging advanced computational approaches through international standards and harmonised guidelines represents a clear way forward, as articulated in the recent "Making eDNA FAIR" paper by Takahashi et al. (2025), which is based on the use of Darwin Core (DwC) and Genomics Standards Consortium (GSC) MIxS standards, as well as Global Biodiversity Information Facility (GBIF)'s "Publishing DNA-derived data through biodiversity data platforms"

* These authors contributed equally to this work.

guidelines. Furthermore, additional complementary principles strengthen this framework. The Collective benefit, Authority to control, Responsibility, Ethics (CARE) principles emphasise Indigenous data governance and responsible sample stewardship, while the Transparency, Responsibility, User focus, Sustainability, Technology (TRUST) principles provide criteria for repository reliability and long-term digital preservation. Together, the combined application of FAIR, CARE, and TRUST principles provides a structured foundation for ensuring robust, interoperable, and ethically managed eDNA data that support aquatic biodiversity research, management, and conservation across Europe.

Key words: Data management, European aquatic biodiversity, metabarcoding, metadata standards, molecular biomonitoring, omics

Introduction

Monitoring biodiversity in aquatic ecosystems is essential for understanding, restoring, managing, and safeguarding the health and resilience of our planet's water environments (Valentini et al. 2016; Simaika et al. 2024). These systems are particularly vulnerable to climate change, habitat degradation, overexploitation, pollution, and invasive species (Corrales et al. 2020; Prakash 2021; Priya et al. 2023; Simaika et al. 2024). Systematic tracking of species diversity and abundance can provide critical insights into ecological health, temporal trends, and emerging threats (Hazen et al. 2019; Fonseca et al. 2023). Such data are foundational for evidence-based conservation and management strategies (Harper et al. 2019; Adams et al. 2024; Altermatt et al. 2025).

A transformative advance in biodiversity monitoring is the application of DNA metabarcoding, a method that identifies multiple species within a mixed environmental sample by sequencing short, taxonomically informative DNA regions. This approach can be applied to bulk samples (e.g., macrozoobenthos) or to environmental DNA (eDNA), which refers to genetic material shed by organisms into their surroundings through skin cells, scales, mucus, faeces, and other biological materials (Leese et al. 2016, 2018; Beng and Corlett 2020; Sahu et al. 2023). The eDNA can be extracted from water, sediment, or other environmental samples with minimal disturbance, enabling the detection of elusive, rare, or invasive species (Jerde et al. 2013; Rourke et al. 2022; Duarte et al. 2023). Its effectiveness for monitoring water quality through indicator taxa has been extensively demonstrated (Cordier et al. 2021). Despite its advantages, interpretation of eDNA-derived data is complicated by environmental and methodological variables, such as DNA degradation under various physical and chemical conditions, which can reduce detectability and bias temporal inferences. Another issue is the transport of DNA particles, which can be decoupled from local organismal presence. As well, the use of multiple, non-standardised protocols has limited the interpretation of results obtained from eDNA sampling (Nukazawa et al. 2018; Stewart 2019). In the remainder of this manuscript, we use the term "eDNA data" to refer collectively to DNA metabarcoding data generated from both environmental samples and bulk organism samples, unless otherwise specified.

As the use of eDNA expands, opportunities to combine datasets across space and time for large-scale biodiversity assessment, synthesis, and long-term monitoring are increasing. Realising this potential, however, depends critically on the infrastructure for storing, sharing, and interoperating eDNA data. Raw sequencing

data are often deposited in repositories like those of the International Nucleotide Sequence Database Collaboration [INSDC: National Centre for Biotechnology Information (NCBI), European Nucleotide Archive (ENA), DNA Data Bank of Japan (DDBJ)], while derived occurrence data may be submitted to biodiversity platforms such as Global Biodiversity Information Facility (GBIF) and Ocean Biodiversity Information System (OBIS). However, the reuse, discoverability, and interoperability of data hosted in these databases depends heavily on the quality and completeness of associated metadata, including sampling methods, sequencing protocols, bioinformatic analysis, and geospatial information (Fig. 1).

The importance of structured metadata in genomic research was recognised early in the development of high-throughput sequencing. In response, the Genomic Standards Consortium (GSC) was established as an open community initiative to develop standards for sequence-associated data (Eloe-Fadrosh et al. 2024). One of its primary outputs, the Minimum Information (MI) about any (x) Sequence (MIxS) standard (Yilmaz et al. 2011), defines a set of metadata elements describing samples and their source environments. MIxS terms related to sampling and sequencing are organised into “checklists”, while environmental descriptors are provided through environment-specific “extensions” (e.g., Air, Soil, Water). For example, the Water extension includes terms such as size_frac_low (mesh or pore size used for pre-filtration) and samp_vol_we_ext (the volume or mass of the sample from which DNA was extracted), both of which are essential for interpreting and comparing aquatic eDNA datasets.

Although such standards aim to improve metadata consistency and comparability, compliance remains variable (Hassenrück et al. 2021). In practice, incomplete adoption or inconsistent implementation of standards limits cross-study integration and large-scale synthesis. To ensure that eDNA data are Findable, Accessible, Interoperable, and Reusable (FAIR), harmonised metadata practices and clearly defined reporting expectations are essential. Recent international initiatives are working to align eDNA sampling and analysis workflows with existing standards to facilitate cross-border comparisons and long-term monitoring (e.g. Takahashi et al. 2025). In parallel, complementary principles such as Collective benefit, Authority to control, Responsibility, Ethics (CARE), which address Indigenous data governance, and Transparency, Responsibility,

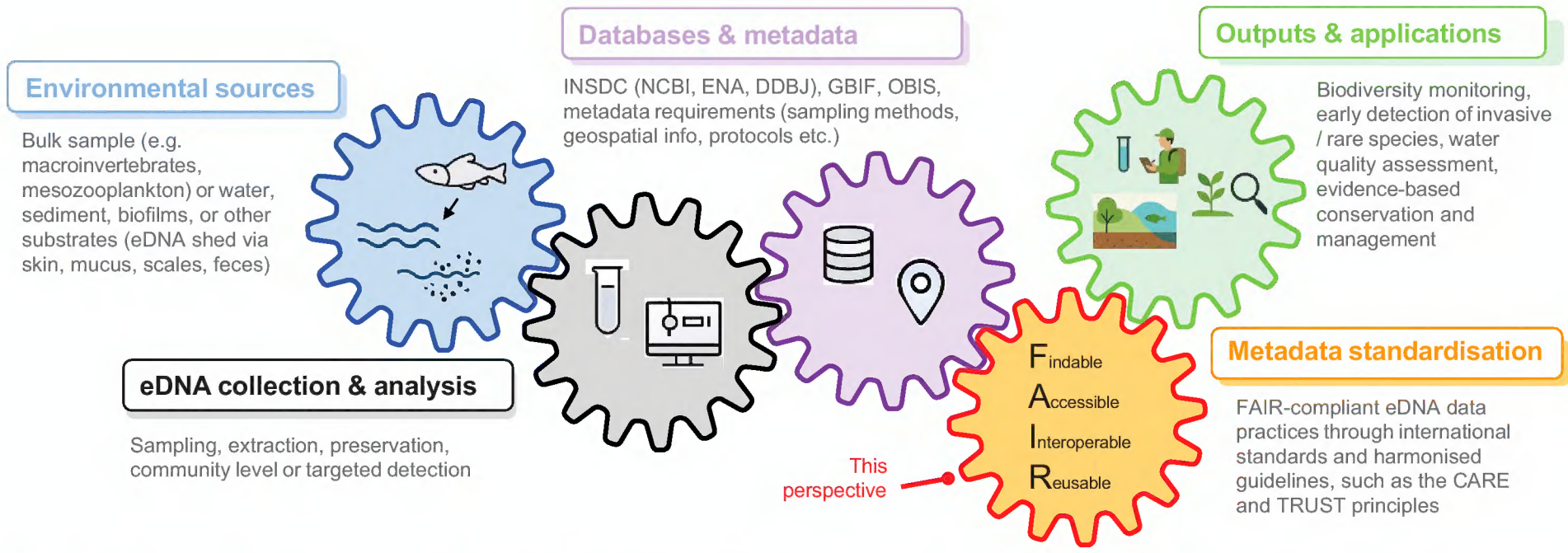
Environmental sources
Bulk sample (e.g. macroinvertebrates, mesozooplankton) or water, sediment, biofilms, or other substrates (eDNA shed via skin, mucus, scales, feces)
Databases & metadata
INSDC (NCBI, ENA, DDBJ), GBIF, OBIS, metadata requirements (sampling methods, geospatial info, protocols etc.)
Outputs & applications
Biodiversity monitoring, early detection of invasive / rare species, water quality assessment, evidence-based conservation and management
eDNA collection & analysis
Sampling, extraction, preservation, community level or targeted detection
Findable
Accessible
Interoperable
Reusable
This perspective
Metadata standardisation
FAIR-compliant eDNA data practices through international standards and harmonised guidelines, such as the CARE and TRUST principles

Figure 1. Conceptual overview of the eDNA workflow, from environmental sampling to biodiversity monitoring outputs, highlighting the central role of metadata standardisation in enabling FAIR-aligned data integration across repositories and applications.

User focus, Sustainability, Technology (TRUST), which guide digital repository stewardship, are gaining traction. Collectively, these frameworks support not only technical interoperability but also ethical, transparent, and sustainable data governance (Lin et al. 2020; Altermatt et al. 2025).

Taken together, these developments highlight a rapidly evolving landscape in which eDNA is increasingly used to inform biodiversity monitoring, management, and policy. However, the pace of methodological innovation, coupled with heterogeneous data storage and metadata practices, risks limiting the effective reuse, integration, and long-term impact of eDNA data, particularly at regional scales where coordinated uptake is essential. To facilitate the best use of eDNA and support its systematic adoption within European aquatic monitoring frameworks, it is therefore timely to take stock of current practices, identify key gaps, and prioritise areas where future effort can deliver maximum impact. This perspective paper contributes to these efforts by building on the Horizon Europe project eDNAqua-Plan ("A Plan towards an eDNA reference library and data repository for Aquatic Organisms, navigating Europe towards the next generation biodiversity monitoring", Grant Agreement No. 101112800). We aim to: (1) evaluate the properties of databases used to store eDNA-derived data, including both raw sequences and taxon occurrences, (2) assess the types, quality, and completeness of metadata submitted by data producers, and determine the extent to which current practices enable data sharing, interoperability, and reuse, (3) identify key challenges, priorities, and opportunities for improving eDNA data standardisation and reuse within the aquatic eDNA community, informed by targeted stakeholder input, and (4) systematically analyse the scientific literature, using a Large Language Model (LLM), to assess how eDNA workflows and associated metadata are documented and reported, and to identify recurring gaps, inconsistencies, and opportunities for improved standardisation.

Methodology

Four complementary approaches were used: 1) a qualitative assessment of publicly available and accessible eDNA databases, in which data formats, accessibility, and their suitability for storing raw sequence data, associated metadata, and derived biodiversity records were evaluated; 2) a quantitative assessment of metadata completeness in eDNA databases; 3) a questionnaire survey on stakeholder eDNA practices; and 4) a LLM-driven analysis of eDNA scientific publications (Table 1). All analyses were conducted within the framework of the eDNAqua-Plan project and were informed by expert teams who provided structured input to identify priority metadata elements and data properties relevant to data FAIRness and to address common questions related to the use and interpretation of eDNA data.

Qualitative assessment of eDNA databases

The eDNA sequencing data are stored in two main ways: (a) primary eDNA databases, which archive raw sequencing reads and annotated sequences (FASTA or FASTQ format); and (b) secondary eDNA databases, which provide processed sequence variants (Operational Taxonomic Unit (OTU)/Amplicon Sequence Variant (ASV) tables) and derived biological identifications, such as taxon occurrence

Table 1. Overview of the four complementary analytical components used to evaluate the European aquatic eDNA landscape, summarising their respective objectives, strengths, and limitations.

Approach	Goals	Strengths	Weaknesses
1. Qualitative Assessment of eDNA Databases	Identify and characterise existing eDNA databases and their key properties.	Provides a structured overview of major eDNA databases, including their scope, geographic and taxonomic coverage, metadata standards, and accessibility features, thereby enabling comparison of repository functionalities and identification of coverage and standardisation gaps.	Dependent on the accessibility of existing databases; may not fully capture internal curation practices or undocumented workflows.
2. Quantitative Assessment of Metadata Completeness in eDNA Databases	Systematic aggregation and evaluation of metadata from databases or repositories.	Enables rapid assessment of data quality and completeness.	Relies on well-structured metadata that are also retrievable via repository interfaces and APIs; does not assess the semantic accuracy or correctness of metadata entries, only their presence or absence; may overlook contextual information reported outside structured metadata fields
3. Questionnaire on Stakeholder eDNA Practices	Gather insights from practitioners on current practices, needs, and challenges.	Provides direct feedback from stakeholders; offers diverse, experience-based perspectives.	Responses may not fully represent the broader community; users may perceive bias or gaps if participation is uneven across sectors or regions.
4. Large Language Model-driven Analysis of eDNA Scientific Publications	Use generative Artificial Intelligence to synthesise and interpret large volumes of scientific literature on eDNA.	Enables rapid identification of patterns, trends, and emerging insights across a large body of scientific publications, substantially reducing the time required for manual literature review.	Outputs may reflect biases or gaps in the model's training data; users must critically validate results to ensure accuracy and reliability.

and biodiversity observations. The eDNAqua-Plan consortium internally identified three major primary eDNA databases belonging to the International Nucleotide Sequence Database Collaboration (INSDC): ENA, DDBJ, and NCBI, and three major secondary eDNA databases, namely GBIF, OBIS, and metaPR2 (Protist Ribosomal Reference database) (Vaulot et al. 2022) (Table 2). It is important to note that these repositories host a wide range of sequencing and biodiversity datasets, the majority of which are not derived from eDNA studies; however, they constitute the primary infrastructures through which eDNA data are currently archived, curated, and disseminated. ENA is a comprehensive repository maintained by the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) that stores raw sequencing reads, assembled sequences, and functional annotations, ensuring long-term preservation and accessibility of genomic data. DDBJ, hosted by the National Institute of Genetics in Japan, archives raw sequencing data, assembled sequences, and annotations, and collaborates closely with ENA and NCBI for data exchange, as a member of INSDC. NCBI, at the National Institutes of Health, is a widely used open-access database based in the United States of America, containing raw sequence data, assembled sequences, and annotations across all taxonomic domains of life. GBIF is an international infrastructure that provides access to species occurrence records from museums, herbaria, and citizen science, and increasingly from molecular sources, including observations derived from eDNA data. OBIS is a database focused on marine biodiversity, integrating species occurrence records, environmental parameters, and increasingly molecular-based observations, including those derived from eDNA. The metaPR2 is a curated database designed explicitly for protist diversity studies, providing taxonomic reference and processed sequence variants from metabarcoding data. In this study, we analysed the type, size, scope, update frequency, usability, and metadata quality (definitions provided in the Suppl. material 1) of each of these databases. Whenever possible, programmatic access was implemented through an Application Programming

Table 2. Qualitative review of six major environmental DNA resources across 40 criteria, summarising their core characteristics and key strengths and weaknesses; all databases (DB) are open access and updated daily, except for metaPR2, which is updated approximately yearly. The finer grained criteria used are shown in Suppl. material 1.

DB	Aquatic eDNA Records (approx.)	Organismal Focus	User Interface	Programmatic Access	Metadata Quality	Strengths	Weaknesses
PRIMARY DATABASES							
ENA	2 m	All	Medium: advanced query builder.	Medium: SWAGGERAPI + fasta files. Standard: MIxS.	Medium: findable & accessible, but low interoperability & reproducibility.	Global and well-resourced for posterity.	Limited curation. Few environmentally useful mandatory metadata terms.
DDBJ	2 m	All	Medium: great for keyword searches.	Medium: limited API + fasta files, but also JSON-LD files. Standard: MIxS.	Medium: findable & accessible, but low interoperability & reproducibility.	Global and well-resourced for posterity.	Limited curation. Few environmentally useful mandatory metadata terms.
NCBI	2 m	All	Medium/ high: easiest for keyword searches.	Medium: custom API + fasta files. Standard: MIxS.	Medium: findable & accessible, but low interoperability & reproducibility.	Global and well-resourced for posterity.	Limited curation. Few environmentally useful mandatory metadata terms.
SECONDARY DATABASES							
OBIS	20 m	Marine organisms	High: intuitive graphics.	Medium: SWAGGER API. Standard: DwC.	High usually.	Tight coupling between OBIS and GBIF. Has key oceanographic properties.	Limited curation. Only marine, no freshwater.
GBIF	2 m	All	Medium/low: Less intuitive.	Medium: API. Standard: DwC.	High usually.	Global. Works with many specialist organisations. Tight coupling between OBIS and GBIF.	Limited curation & amount of eDNA-derived data.
metaPR2	2,000 (samples)	Protists	Medium: R-shiny interface.	Low: download files or use Docker). Standard: Non-applicable.	High: extensive curation.	Highly curated.	Has no API.

Interface (API) to retrieve the required data. General-purpose data repositories (e.g. Dryad) were not included, as they primarily function as archival platforms and do not provide standardised eDNA-specific data models, taxonomic integration, or biodiversity indexing required for large-scale interoperability and reuse.

Quantitative assessment of metadata completeness in eDNA databases

The MIxS checklists and extensions define standardised metadata terms for INSDC submissions, classified as mandatory, recommended, or optional. Mandatory metadata for biological samples include sample name, collection date, environmental context (major and local environmental system), geographic origin, and geographic coordinates (latitude and longitude) (Suppl. material 1). Recommended terms encompass details on sampling procedures, DNA isolation, PCR amplification, and sequencing. The MIxS “Water” extension additionally contains metadata fields describing environmental parameters such as alkalinity, temperature, and nutrient concentrations. Our analysis focused on metadata describing geographic location and environmental ontology, as well as properties intrinsic to the sample and associated sequence data, including sampling protocol, nucleic acid extraction and amplification details, target gene or its amplified fragment, primer sequences, and sequencing method.

Information about high-throughput sequencing data in the INSDC is stored within a hierarchical structure that can broadly be described as: Study/BioProject > Sample/BioSample > Experiment. The BioSample (referred to as "Sample" in ENA) is intended to hold descriptive information about the biological source material, whereas the Experiment metadata describes the methodology associated with a specific sequencing run. To assess the richness and machine-readability of metadata for European aquatic eDNA sequence data, we retrieved Sample/BioSample accession numbers using an Entrez search with the following terms: ("eDNA"[All Fields] OR "environmental DNA"[All Fields]) AND ERC000024[All Fields], where ERC000024 corresponds to the GSC MIxS Water checklist in ENA. This search strategy retrieved samples explicitly submitted under the MIxS Water checklist and described by submitters as eDNA. While some eDNA datasets may not explicitly use these terms (e.g., samples submitted as marine or freshwater metagenome), this approach ensured consistency by limiting results to studies in which authors defined their samples as eDNA. Most retrieved records were submitted within the past 5–10 years (data not shown), thereby reflecting more recent submission practices. Custom Python scripts were used to retrieve all metadata associated with each Sample accession number, as well as all linked Experiment records and their associated metadata. Sample entries without associated Experiments were excluded from further analysis. All retrieved metadata fields were compiled, harmonised, and categorised for subsequent evaluation.

The Environmental Ontology (ENVO) provides a structured vocabulary for describing environmental entities (Buttigieg et al. 2013) and is recommended within MIxS to standardise environmental descriptors. We quantified the use of ENVO terms by calculating the proportion of samples containing ENVO identifiers in the mandatory metadata fields env_medium, env_local_scale, and env_broad_scale. ENVO usage was identified by searching for the string "envo" (case-insensitive) within these fields. For other metadata categories, the threshold for "reusability" can be context-dependent. Nevertheless, we targeted a set of metadata fields considered essential for comparative analyses and meta-studies. These included metadata related to sample collection, size fractionation, DNA isolation, primer sequences, PCR amplification, and sequencing strategy. Data processing and statistical summaries were conducted in R (v4.3.2; R Core Team 2023).

To complement the analysis of primary sequence repositories, we evaluated metadata completeness in a secondary biodiversity data aggregator. The OBIS was used as an example platform because it exclusively hosts marine biodiversity records. DNA-derived records were queried using the OBIS "mapper" interface, and European datasets were extracted using a spatial polygon filter. The availability and consistency of environmental, sampling, and molecular metadata fields associated with these records were assessed using the same criteria described above.

Questionnaire on researchers' eDNA practices

A structured questionnaire was developed to document current eDNA research practices and to identify patterns of uniformity, variability, and alignment with community standards across the complete eDNA workflow, from sampling to data archiving (35 questions; Suppl. material 1). The survey addressed aspects of data generation, including project scope, geographic coverage, sampling locations, ecosystem types, substrates, genetic markers, primers, sequencing

platforms, and approximate sample and read numbers. Questions on data management addressed whether raw and processed reads were deposited, the types of repositories used (e.g., NCBI, ENA, OBIS, Zenodo, GitHub), accessibility status, and the circumstances that affected deposition. Metadata practices were explored through questions on the types of metadata compiled (e.g., geographic coordinates, sampling protocols, DNA extraction and sequencing methods, bioinformatics pipelines, and environmental data), the adoption of community standards (e.g., Darwin Core, MIxS, Minimum Information about a marker gene sequence (MIMARKS)), and whether metadata were archived alongside eDNA data. Finally, the questionnaire examined repository interoperability by asking whether deposited eDNA datasets were linked to other repositories and, if so, which ones. The survey was initially circulated within the eDNAqua-Plan consortium and subsequently distributed more broadly to researchers working on European aquatic biodiversity via the eDNAqua-Plan and the iBOL-Europe networks.

LLM-driven analysis of eDNA scientific publications

To examine methodologies applied in eDNA research, an initial set of 117 relevant scientific publications was assembled, manually curated, and reviewed by subject-matter experts. Each publication was systematically annotated according to predefined methodological criteria: (1) whether the reported data are environmental in nature; (2) whether data are accessible in the article or Suppl. material 1; (3) the sample collection method; (4) the DNA extraction method; (5) whether the protocol is deposited in protocols.io (other protocols DOIs were not surveyed); (6) the sequencing strategy employed; (7) the sequence analysis workflow; (8) the data repository used for sequence deposition; (9) the genetic marker analysed; and (10) the reference database applied for taxonomic identification (available at: GitHub repository). This curated corpus provided a high-quality training dataset, which was subsequently leveraged to automatically identify and incorporate additional relevant literature through the LLM, ultimately expanding the collection to 1,607 publications.

An LLM was used to systematically extract methodological information from the expanded corpus of publications identified beyond the manually curated training set. The pre-trained Llama 2 model was further fine-tuned to focus on typical elements commonly found in eDNA studies, particularly those described in the Materials and Methods sections. The use of the LLM enabled the identification of relevant patterns across large volumes of text, including protocols for sample collection, DNA extraction, sequencing strategies, and data repository practices. The model's outputs were manually reviewed to ensure consistency with expert-defined categories and were used to automatically generate structured summaries of methodological trends across the literature. Further methodology and validation practices are detailed in Suppl. material 1.

To comprehensively analyse trends within eDNA research, we also employed Medical Subject Headings (MeSH) terms obtained via the PubMed Entrez API (Sayers 2009). A custom program was developed to query the API using the LLM-extracted Digital Object Identifiers (DOI) (available at: GitHub repository). MeSH terms represent a controlled vocabulary developed and maintained by the National Library of Medicine, used for indexing biomedical literature in PubMed (Dhammi and Kumar 2014). MeSH terms provide well-defined, curated descriptors that ensure

consistency across publications. This standardisation is particularly valuable when examining research trends over extended time periods, as it eliminates variations in terminology that could obscure underlying patterns (Yang and Lee 2018).

Results

Qualitative assessment of eDNA databases

Our qualitative review of six major eDNA databases revealed significant differences in their use of metadata standards, accessibility, and curation practices (Table 2). The INSDC sequence databases (ENA, DDBJ, and NCBI) are broad and global in scope, each hosting approximately two million aquatic eDNA sample-level records, with new submissions added daily. Focusing on Europe, Fig. 2A shows the total number of aquatic eDNA sequence records available in INSDC as of June 2024, while Fig. 2B presents the corresponding number of

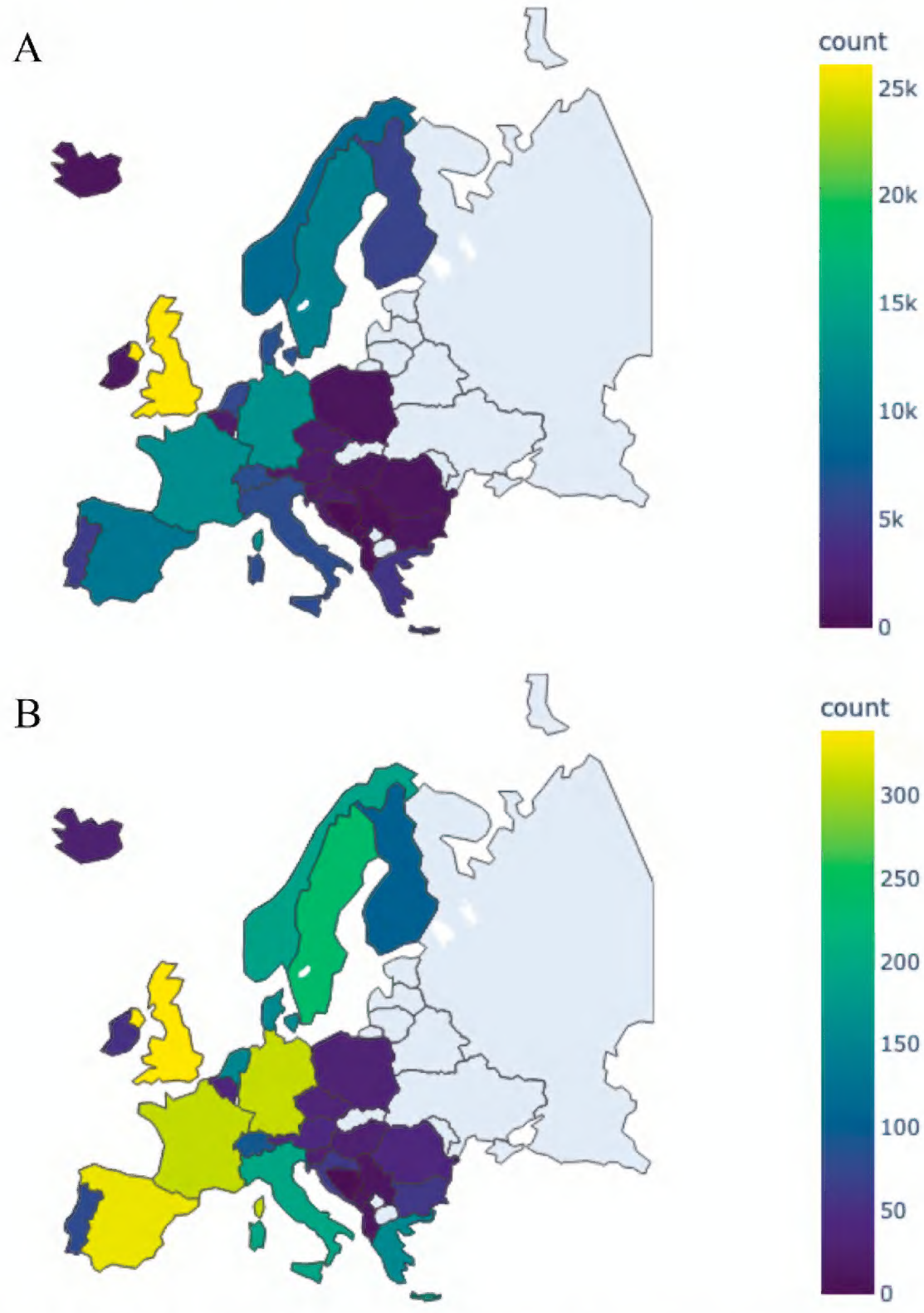

Figure 2. Counts of aquatic environmental DNA sequence records from Europe available in the International Nucleotide Sequence Databases Consortium (INSDC), as of June 2024. **A.** Number of sequence records; **B.** Number of projects/studies for which a sequence has been submitted. White indicates water bodies, and light blue indicates areas for which data were not available.

BioProjects/studies for which a sequence has been submitted. These repositories are characterised by long-term institutional support, open access policies, and mature human-oriented search interfaces. While MIxS-compliant environment-related metadata fields are supported, their population is heterogeneous across records, and several environmentally relevant attributes are optional rather than mandatory, resulting in variable metadata completeness across submissions, as further documented below. By contrast, OBIS and GBIF are explicitly designed for biodiversity data integration, leveraging the Darwin Core (DwC) standard to achieve higher metadata resolution. Both are tightly coupled, updated daily, and provide APIs that facilitate machine access, though OBIS is restricted to marine data, whereas GBIF spans all taxa across all biomes. Finally, metaPR2 is the most specialised resource, focusing on protists, with around 2,000 highly curated samples. While it excels in metadata quality, its usability for machine access is limited due to the absence of an API. Overall, the comparison highlights systematic differences among database types that reflect their original scope and design objectives: primary sequence repositories prioritise comprehensive archival of raw sequence data at a global scale, whereas biodiversity-focused secondary resources emphasise harmonised ecological and occurrence metadata structured for integration and synthesis across studies. Specialised resources such as metaPR2 further illustrate how targeted curation can enhance metadata quality within specific taxonomic groups.

Quantitative assessment of metadata completeness in eDNA databases

Primary eDNA repositories

We recovered 5,280 sample accessions with the Entrez search, excluding entries without any associated experiments. Of these, 82% were submitted after 2020. In total, 130 unique metadata fields were used to describe the retrieved samples. Of these, 16 were used to describe the geography or environment (Suppl. material 1), but the terminology was somewhat inconsistent; for example, local environment was denoted either as “local environmental context” or as “env_local_scale”, whereas “depth”, defined as the vertical distance below local surface, was also denoted “geographic location (depth)”. Similarly, oceanic region was submitted under either “geographic location (ocean)” or “geographic location (region and locality)”. Latitude and longitude were in most cases given as decimal degrees, for fewer than 10 samples they were given as Degrees-Minutes-Seconds. All of the samples had metadata under the mandatory fields “env_broad_scale”, “env_local_scale” and “env_medium”, however, Environment Ontology (ENVO) terms were used in only about 18% of the entries for each of these.

With regards to sampling, laboratory processing, and sequencing, we identified several terms which could be associated with each step (Suppl. material 1). We investigated more closely the terms listed in Table 3. Generally, the submitters provided little information on sampling and lab work, as for instance sample collection method and size fraction was provided in less than 20% of the records (Table 3). Filter pore size was also sometimes provided as free text under sample collection method or device. The volume or mass of sample used for DNA extraction

Table 3. Use of recommended MIxS+Water terms in metadata for records described as "environmental DNA" or "eDNA" in ENA, or datasets with the DNADerived extension from European waters in OBIS. n = 5,283 BioSamples and 1,242,198 occurrences (corresponding to 25 data sets), respectively.

Sample metadata term	% with information, ENA	% with information, OBIS	Example of entries
samp_collect_device/'sample collection device'/samp_collect_method	2%/18%/12%	0%	Text strings with names and types of filters including pore size, brand name of water sampler
size_frac_low/size_frac	13.5%	0%	0,0008, 0,0012, 0,1, 0.0012, 0.2, 3
size_frac_up	1.6%	0%	0,005, 200, 3
samp_mat_process	0.1%	0%	Reference to internal protocol
samp_store_loc, samp_store_temp	0.1%, 3%	Not used	«eDNA freezer bag», -80, -20, 4, 20
samp_vol_we_dna_ext	1.5%	59%	Numerical values
nucl_acid_ext/'nucleic acid extraction'	2.2%	59%	Links to protocols.io, reference to papers
pcr_conditions, nucl_acid_amp	2.4%	59%	PCR protocol as text string, reference to papers
pcr_primers/ pcr_primer_*	23%	99%	ENA: sequences or names of PCR primers, OBIS: Sequences of PCR primers
target_gene, target_subfragment	35%, 2%	90%, 74%	Names of target gene
seq_meth	52%	91%	Name of sequencing platform
otu_class_appr, otu_db, otu_seq_comp_appr	Not relevant	60%	Name of bioinformatic pipeline, name of reference database. Metadata field tax_class was used for 2% of records, where name of ref db and other parameters were given.

was reported in the field "samp_vol_we_dna_ext" in 1.5% of cases, while information on nucleic acid extraction methods under "nucleic acid extraction" appeared in 2%. The MIxS term "nucl_acid_ext" did not appear in any of the records. Multiple amplifications or library preparation methods may be performed on the same eDNA sample, thus it would be logical to store metadata such as PCR primers, target gene and gene fragment, and PCR conditions as experiment metadata (see Takahashi et al. 2025 for alternative data structures). Nevertheless, this information, when provided, was in most cases stored as sample metadata. Target gene was given for 35% of the records, whereas the target gene fragment was provided for only 2%. PCR primers were given for 23% of the records, either as the actual sequences, or names of primers. In a few cases the terms "gene_target" (24 records) and "library_design" (118 records, ~2%) were used as experiment attribute tags. Information on primers and target gene was also sometimes submitted as free text in the "description" field for each "Sample". Library selection, library source, and library strategy, as defined in the ENA Experiment checklist, were provided for all records (Suppl. material 1).

Secondary eDNA repositories

Analysis of OBIS identified approximately 110 datasets associated with the DNA-derived data extension, of which 25 included sampling locations in Europe (last accessed January 2026). Across these datasets, 143 metadata fields from the DNA-derived extension were populated by at least one dataset. In addition, the associated occurrence metadata comprised approximately 280 fields, including taxonomic identifiers and related attributes.

The DwC DNA-derived data extension provides structured fields describing elements of the analytical workflow, including sample volume, laboratory protocols, sequencing methods, OTU clustering approaches, reference databases, and annotation algorithms. For instance, the field "otu_class_appr", designed to hold technical metadata about OTU clustering, was completed for 60% of the

occurrences in the OBIS European records (Table 3). However, specific fields for documenting sequence denoising methods or explicitly quantifying uncertainty in taxonomic assignments are currently lacking. In several cases, such information was instead provided as free text within general-purpose fields such as “identificationRemarks” in the occurrence metadata. For example, entries included statements such as “rdp annotation confidence: 0.8; blast coverage threshold of 75% and identity threshold of 97%” or “RDP classifier; confidence cutoff 0.8; ref db Midori v2.” The “identificationReferences” field occasionally contained links to GitHub repositories, online protocol repositories, GenBank records, or publications describing specialised taxonomic reference lists.

Questionnaire on researchers’ eDNA practices

A total of 69 respondents from 29 countries participated in the questionnaire survey, representing 42 academic institutions, 19 governmental institutions, two intergovernmental organisations, two research centres or infrastructures, two non-governmental organizations, one foundation, and one private company. Responses reflected a strong emphasis on marine and freshwater environments, while transitional ecosystems were less frequently reported (Fig. 3A). The most commonly used genetic markers were the universal metabarcoding loci, such as the mitochondrial 16S, 18S, COI, and nuclear ITS, whereas more taxonomically focused markers (e.g., 12S, rbcL, matK, tufA) were used less consistently across studies (Fig. 3B). Most respondents reported depositing their data in primary sequence repositories, particularly GenBank/NCBI/ENA, often in combination with Barcode of Life Data System (BOLD). Use of general-purpose or custom platforms (e.g., Zenodo, Phytool,

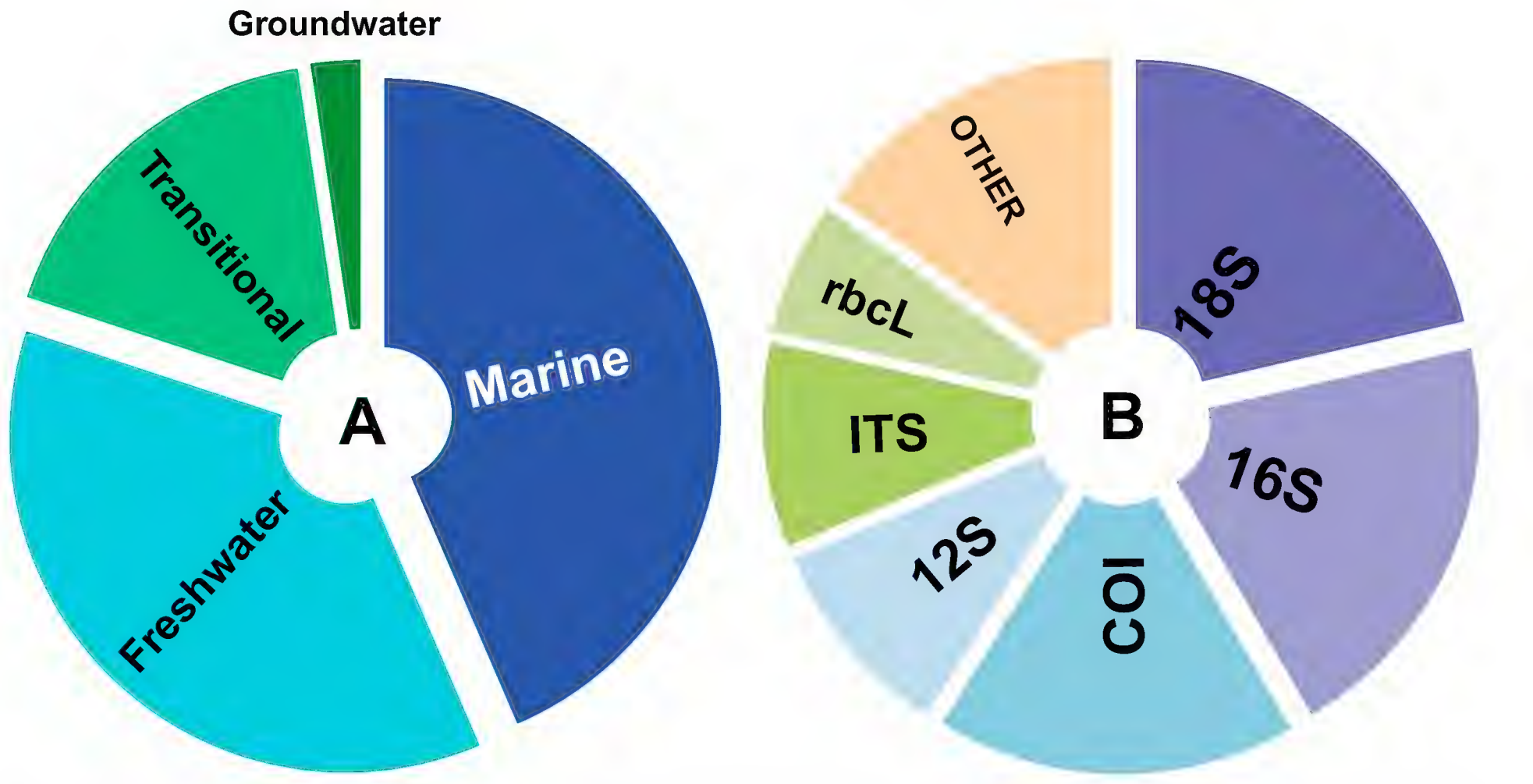

Figure 3. Results of the questionnaire survey on aquatic eDNA practices. **A.** Distribution of aquatic environment types targeted by respondents: Marine (43.5%), Freshwater (36.5%), Transitional (17.6%), and Groundwater (2.4%); **B.** Genetic markers reported as most frequently used for eDNA metabarcoding: 18S (21.4%), 16S (20.2%), COI (17.3%), 12S (9.8%), ITS1–5.8S–ITS2 (9.8%), rbcL (6.4%), and other markers (15.1%), which included 23S, 28S, cytb, metatranscriptome, metagenome.

MIMt, MetaZooGene, institutional or project-specific databases) was limited to a few cases. Only a small fraction of respondents indicated that their data had not yet been deposited or were awaiting inclusion in forthcoming repositories. Sequencing platforms were dominated by Illumina technologies, with MiSeq (47 mentions) being the most widely used, followed by HiSeq (19), NovaSeq (16), and NextSeq (8). Other platforms, such as Nanopore (10), the now discontinued Roche 454 (7), PacBio Sequel (4), and Ion Torrent (across four models), were less frequently used. Metadata reporting practices varied considerably: in primary repositories, sampling methods were reported by 59% of respondents and DNA extraction protocols by 51%, whereas sequencing platforms were included by 56% but PCR protocols by only 2%. Similar trends were observed for secondary repositories, where approximately half of respondents reported sampling and extraction metadata, 46% documented sequencing platforms, and none reported PCR protocols. Eleven out of 69 respondents reported submitting eDNA derived data to secondary databases/biodiversity portals such as GBIF, OBIS, and metaPR2. Fewer than 10 reported submitting processed data to more general repositories such as Zenodo and FigShare or publishing it in the Suppl. material 1 of the publication.

Analysis of LLM-driven literature review of eDNA metabarcoding studies

The LLM-driven literature collection resulted in a curated database of 1,607 eDNA publications in total, including the 117 manually curated publications used for model training and validation, spanning multiple years, research fields, and methodological approaches, thereby providing a comprehensive overview of the eDNA landscape. This resource was compiled into a dataset containing the complete set of LLM-generated outputs for all collected publications (available at: GitHub repository). Each database entry includes the publication text organised by section, structured metadata with bibliographic and reference details, and the corresponding LLM-generated answers to our predefined question set.

The retrained LLM model performed well in extracting methodological information from scientific text, with validation showing an accuracy of 97–99% when compared against entries in the training set. Summaries of methodological details were generated across partially overlapping subsets of eDNA papers, as each research question was addressed independently, and extraction success depended on the explicit availability of information within each publication. Consequently, not all papers provided sufficient detail to answer every question. Out of the 1,607 publications analysed, data documentation practices could be inferred for 1,440 papers. Within this subset, 64% reported experimental data in Suppl. material 1, 32% embedded data within the main text, and fewer than 5% used both approaches. With respect to sample collection and laboratory methodologies, the analysed publications encompassed a wide range of sampling and experimental approaches, reflecting the diversity of ecological contexts and study designs within eDNA research. The LLM was able to consistently identify and categorise these broad stages across publications, although the level of methodological detail reported varied substantially among studies.

Out of the 1,607 papers analysed, the LLM was able to extract or infer information on data storage repositories from 731 entries, providing insight

into current documentation and reporting practices in aquatic eDNA research. GenBank was by far the most commonly cited repository, appearing in 711 papers (97.3% of those with identifiable repository information). In contrast, references to alternative repositories were comparatively rare and fragmented, including isolated mentions of the BOLD, FigShare, GitHub, and Mendeley Data, as well as occasional thematic or region-specific infrastructures. Importantly, the LLM analysis revealed substantial inconsistencies in how methodological documentation is reported. Of the 1,440 publications for which methodological information could be inferred, only six (0.4%) cited experimental protocols deposited in protocols.io. Moreover, none of the associated protocol DOIs were included in the formal reference sections of the respective papers, suggesting weak integration of externally hosted protocols into standard citation practices. This pattern highlights a broader gap in transparent and standardised documentation of laboratory workflows, despite the availability of FAIR-compliant platforms. Other infrastructures that enable version-controlled and citable protocol sharing, such as GitHub with DOI minting via Zenodo, also existed but were not systematically assessed in this analysis. Beyond repository use, the LLM-derived synthesis further demonstrated a strong concentration of studies around a limited set of genetic markers. Among the 867 publications in which markers could be identified, the standard metazoan barcoding marker (mitochondrial COI) was most frequently reported (19.0%), followed by 16S rRNA (17.0%), and ITS (14.5%). Other markers, including trnL, ITS1, and SSU, were less common, while LSU and additional targets each accounted for fewer than 1% of studies (Fig. 4A). This concentration reflects methodological convergence but

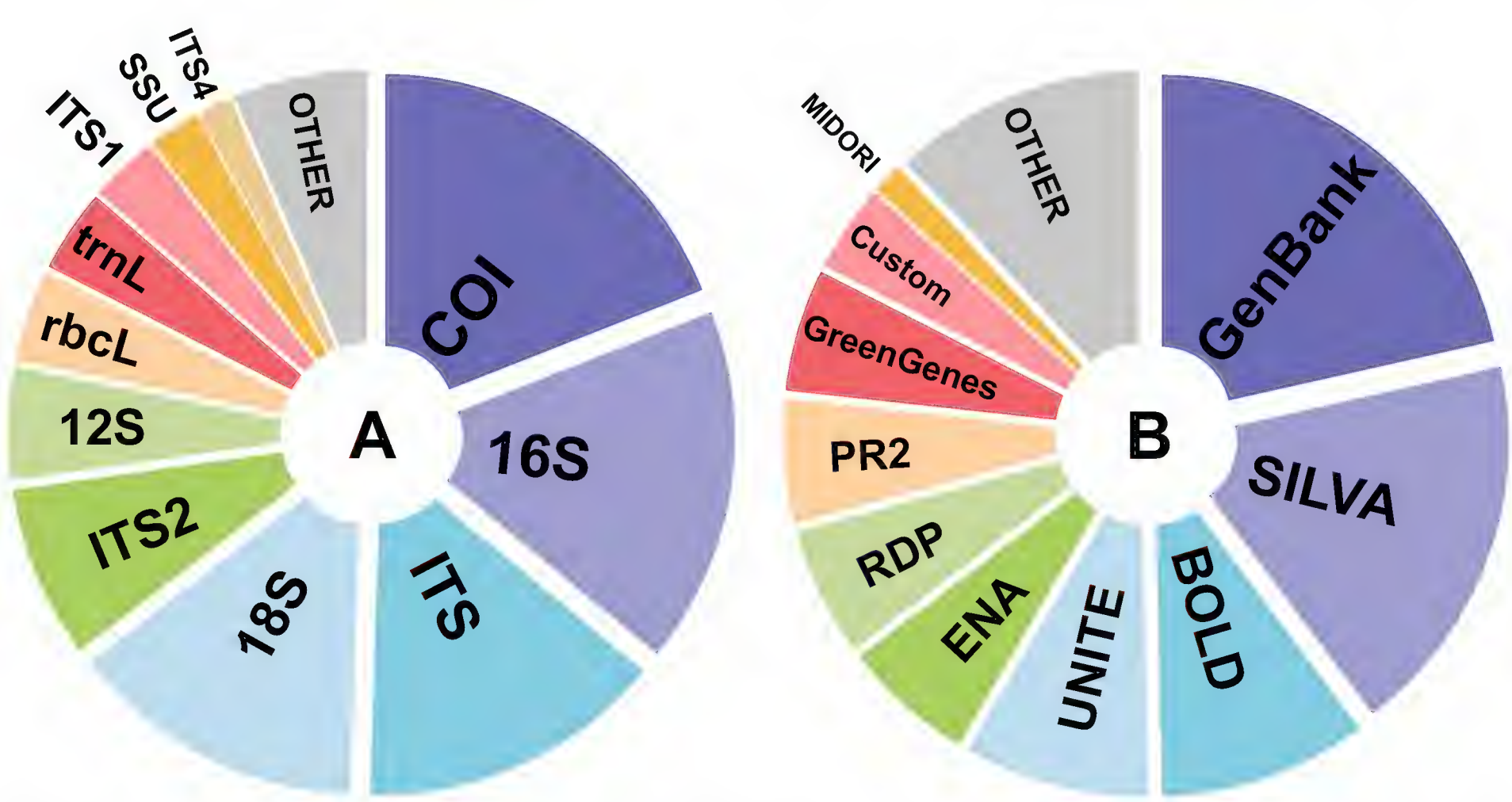

Figure 4. Reference markers and databases used in eDNA studies. **A.** Frequency of genetic markers reported across the 1,607 eDNA papers analysed with the Large Language Model (LLM). The most common markers were COI (n = 165), 16S (147), ITS (126), and 18S (122); **B.** Frequency of reference databases cited for taxonomic assignment. The three dominant resources were GenBank (n = 193), SILVA (164), and BOLD (89). Markers and reference databases concerning less than 1% of entries are grouped under 'other'. Complete counts are provided in the Suppl. material 1.

may also constrain cross-study comparability where reporting standards differ. Collectively, these findings illustrate not only repository preferences and marker trends, but also structural weaknesses in how eDNA workflows and documentation are reported and linked, underscoring the need for improved integration of FAIR-compliant protocol citation and metadata practices.

Concerning taxonomical identification, three major databases dominated usage as reference datasets. GenBank was the most frequently cited, appearing in 19.7% of 898 papers where the model could infer a reference database. SILVA followed closely at 18.3%, reflecting its widespread adoption in studies based on the ribosomal RNA gene. BOLD accounted for 9.9% of responses, due to its relevance in COI-based barcoding applications. Other commonly reported databases included UNITE (8.8%), ENA (6.2%), RDP (6.1%; Ribosomal Database Project, focused on curated bacterial and archaeal rRNA sequences), PR2 (5.9%; Protist Ribosomal Reference database, dedicated to eukaryotic microbes), and Greengenes (5.8%). Less frequently mentioned databases, such as MIDORI, MitoFish, and DDBJ, each contributed under 2% of responses. Notably, custom-built reference databases were reported in 4.1% of studies (Fig. 4B).

Discussion

Structural challenges in the expanding eDNA landscape

Applying eDNA to biodiversity research and monitoring in aquatic ecosystems is proving transformative for conservation and ecological assessment. The benefits of eDNA techniques, notably non-invasive sampling and comprehensive taxonomic coverage, are extensively documented and widely recognised. However, as the field matures, several structural challenges are becoming increasingly apparent. These include persistent gaps in metadata standardisation and completeness, limited interoperability across repositories and standards, discrepancies between reported and actual data practices, regional inequities in data representation, and inconsistent methodological and indexing practices that constrain reproducibility and large-scale synthesis. Addressing these interconnected issues is essential if eDNA data are to become fully interoperable, machine-actionable, and reusable at a global scale.

Publishing eDNA data and choice of database

Our results highlight inconsistencies in metadata reporting, discrepancies between reported and actual data submissions, and substantial regional variation in eDNA application. Analysis of European aquatic eDNA data archived in INSDC reveals pronounced geographic asymmetries (Fig. 2). Fig. 2A shows that total sequence record counts are heavily concentrated in Northern and Western European countries, whereas Eastern and Southeastern Europe contribute substantially fewer records. Fig. 2B demonstrates a similar pattern in the number of BioProjects/studies, indicating that the imbalance reflects not only differences in sequencing volume but also disparities in project-level participation. Large discrepancies indicate that in some regions DNA research activity is present but projects may be smaller in scale, generate fewer samples per study, or involve lower sequencing

intensity. Conversely, countries with both high project counts and high sequence volumes likely host larger-scale or more deeply sequenced initiatives.

Beyond geographic disparities in data deposition, we observed substantial shortcomings in reporting transparency. The inability to reliably identify deposition repositories in more than half of the publications underscores substantial gaps in transparent and standardised documentation. Among the subset of papers for which deposition information could be inferred (n = 731), 97.3% deposited their data in GenBank, demonstrating a strong reliance on INSDC repositories. These findings were consistent with responses from the questionnaire survey, in which most participants reported deposition of raw sequencing data in GenBank or ENA. In contrast, the use of general-purpose or custom repositories (e.g., Zenodo, FigShare) was limited. Our qualitative evaluation of the INSDC databases suggests that their strengths lie in long-term institutional sustainability, open-access policies, and well-developed human-oriented search interfaces, making them trusted archival platforms for primary sequence data. Compared with general-purpose repositories, they provide more structured access to sequencing metadata. However, inconsistencies in INSDC, APIs, and search functionalities limit fully reliable and scalable machine access. For instance, none of the INSDC repositories offered straightforward retrieval of comprehensive, aquatic-focused records without complex query construction.

The submission of taxonomically assigned, eDNA-derived occurrence data to biodiversity platforms (i.e., secondary databases) appears to be less common and remains relatively recent. At the time of writing, we recovered 25 datasets in OBIS that contained the DNA-derived extension from European marine waters. The questionnaire survey similarly indicated that processed sequence data submission to biodiversity portals has lower priority among researchers, with only 15% reporting such practices. A comparable proportion reported submission of processed data or occurrence tables to general repositories such as FigShare or Zenodo.

Together, these findings suggest that while primary sequence deposition in INSDC repositories is well established, the integration of processed eDNA-derived biodiversity data into secondary platforms remains limited. Strengthening awareness of the added value of biodiversity portals for ecological synthesis, interoperability, and policy uptake may help promote more systematic submission of derived data products alongside raw sequences.

Metadata in databases: Insights from automated assessment and self-reporting

Although international guidelines exist and progress is being made, eDNA field names and laboratory methodologies remain insufficiently standardised, leading to inconsistencies in recorded data and challenges for interoperability and comparability. These findings indicate that current reporting practices remain heterogeneous across studies and repositories. In particular, inconsistencies in terminology and environmental descriptors complicate cross-study synthesis and automated data integration. The lack of environmental metadata standardisation further restricts cross-study comparisons. Even where standards exist, they are not always compatible:

the GSC MIxS framework is used by INSDC, while DwC underpins GBIF and OBIS. Encouragingly, recent efforts have begun to improve interoperability between these frameworks. The core terms of MIxS and DwC have been mapped (Meyer et al. 2023), and both standards are incorporated within the FAIR eDNA (FAIRe) guidelines (Takahashi et al. 2025), which aim to support consistent and machine-actionable metadata across repositories. In addition, ongoing work toward deeper harmonisation of FAIRe, MIxS, and DwC (Takahashi et al. 2026) further strengthens the prospects for coordinated implementation across biodiversity and sequence data infrastructures. Takahashi et al. (2025) also provide a structured metadata checklist and formatting guidelines specifically tailored to eDNA workflows, drawing from both MIxS and DwC to clarify which contextual, laboratory, and bioinformatic metadata elements should accompany eDNA datasets in repositories. The quality and completeness of metadata in INSDC repositories were highly variable. For instance, target gene (=barcode) nomenclature was often inconsistent, and fields such as “broad-scale environmental context” and “local environmental context”, both intended to use ENVO ontology terms, frequently contained unstandardised values. In practice, harmonising such information requires the development of complex, regular expressions and considerable manual curation, highlighting that much of the metadata remains insufficiently standardised for automated integration and therefore does not yet meet the expectations of machine-readable FAIR data.

When comparing questionnaire responses to the automated metadata survey, we noted that while respondents reported higher compliance with metadata submission practices, actual analysis of metadata from INSDC BioSamples revealed a lower-than-expected volume of reported metadata. This discrepancy suggests a significant gap between reported intentions and actual practices. One explanation for this gap may lie in survey-related biases, including self-selection of more engaged or experienced respondents and social desirability bias, both of which could lead to overreporting of good metadata practices. Differences in interpretation of what constitutes “complete” metadata submission may also have contributed to mismatches between perceived and repository-recorded compliance. Beyond explaining the discrepancy itself, several structural factors likely contribute to the generally low level of metadata completeness observed in repositories. These include misperceptions of the full scope of metadata requirements, lack of training or awareness, the complexity of submission systems, time and resource constraints, and limited incentives or enforcement. Although our automated metadata inventory specifically targeted records submitted under the MIxS Water checklist, the observed heterogeneity in metadata reporting suggests that checklist selection and implementation alone do not guarantee comprehensive metadata submission. More generally, where habitat- or domain-specific checklists are not fully utilised or completed, submitters may not be sufficiently guided to provide environmentally relevant descriptors. Such variability in checklist use and completion may negatively impact the findability, interoperability, and reusability of data by limiting the ability to evaluate data reliability or reproducibility. This creates a reinforcing cycle in which incomplete or low-quality metadata reduces dataset trustworthiness, discourages reuse, and further weakens incentives for comprehensive reporting.

Standardisation as a structural requirement for FAIR interoperability

Across our analyses, the most recurrent limitations, that is, heterogeneous terminology, inconsistent checklist usage, incomplete contextual metadata, and the need for manual harmonisation, point to a common underlying issue: the absence of consistently applied and operationalised reporting standards across repositories and studies. Our findings therefore highlight the need for standardised metadata frameworks that will improve the accessibility, interoperability, and reusability of eDNA data in line with FAIR principles (Fig. 5).

In particular, environmental descriptors were frequently inconsistently filled, and barcode nomenclature varied across records, reducing machine-actionability and complicating cross-dataset integration. Although our study did not explicitly compare metadata completeness between habitat-specific and non-habitat-specific checklist submissions, the observed variability in environmental and methodological fields indicates that checklist selection and completion practices remain uneven. These patterns suggest that the low and mixed practices documented here are not solely behavioural, but are also structural, reflecting gaps in guidance, enforcement, and harmonised reporting expectations. Findability can be enhanced by promoting data deposition in sequence repositories and by strengthening links between eDNA repositories and biodiversity databases such as GBIF and OBIS (Abarenkov et al. 2024), thereby enabling broader access for researchers, environmental managers, and policymakers. Metadata completeness and adherence to standards are crucial for effective data governance under the CARE and TRUST principles (Lin et al. 2020). While good metadata standards form a critical foundation, true interoperability also requires standardised formats and ontologies, machine-readability, persistent identifiers (PIDs), and the technical

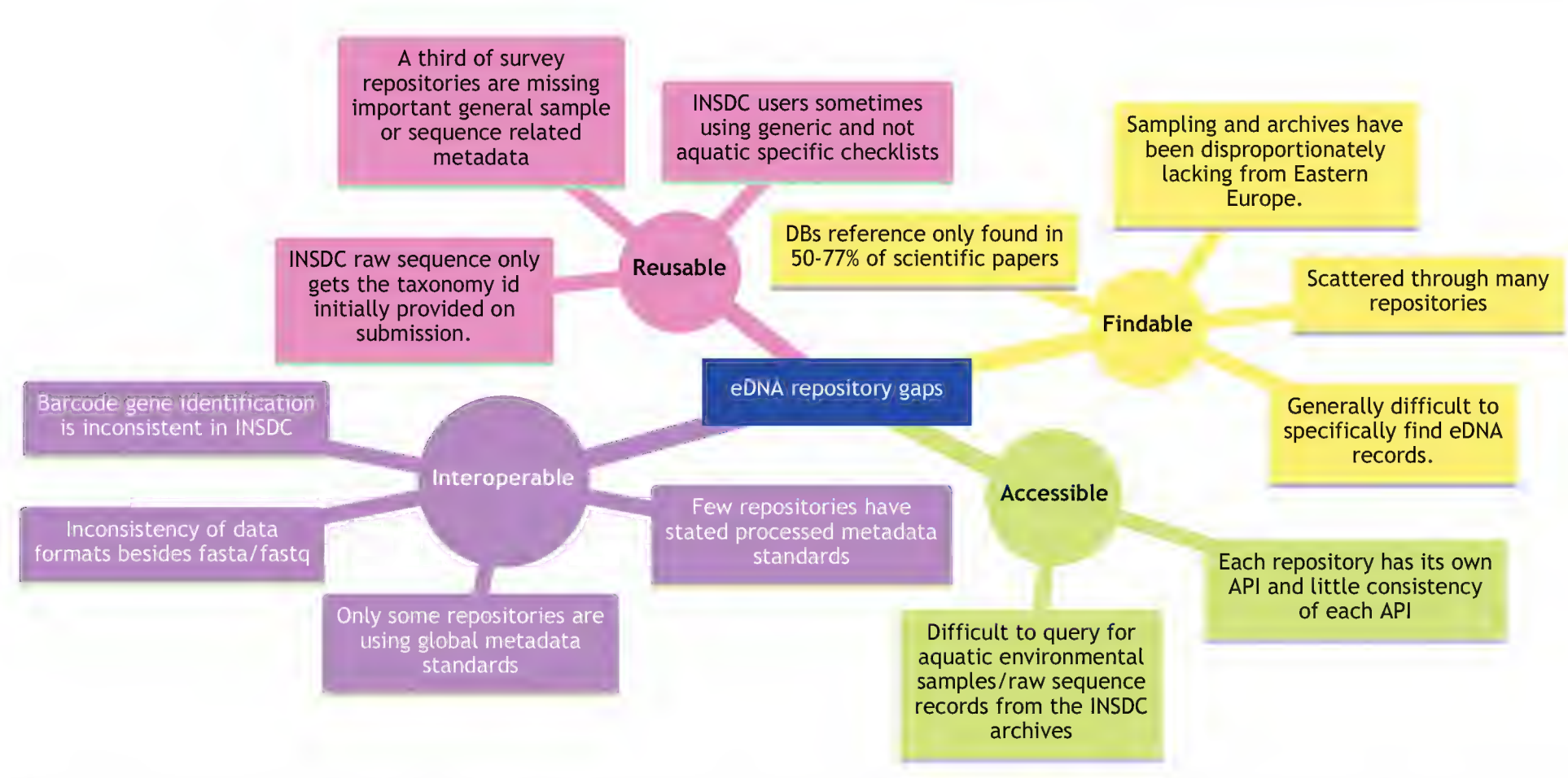
A third of survey repositories are missing important general sample or sequence related metadata
INSDC users sometimes using generic and not aquatic specific checklists
Sampling and archives have been disproportionately lacking from Eastern Europe.
INSDC raw sequence only gets the taxonomy id initially provided on submission.
Reusable
DBs reference only found in 50-77% of scientific papers
Findable
Scattered through many repositories
Generally difficult to specifically find eDNA records.
eDNA repository gaps
Barcode gene identification is inconsistent in INSDC
Interoperable
Inconsistency of data formats besides fasta/fastq
Few repositories have stated processed metadata standards
Only some repositories are using global metadata standards
Accessible
Each repository has its own API and little consistency of each API
Difficult to query for aquatic environmental samples/raw sequence records from the INSDC archives

Figure 5. Summary of the gaps across the spectrum of Findable, Accessible, Interoperable, and Reusable (FAIR) principles in environmental DNA (eDNA) repositories.

infrastructure to support seamless data exchange across repositories. These measures ensure datasets are interoperable with others and remain reusable over the long term. International frameworks such as CoreTrustSeal, DIN31644/NESTOR, and ISO163638 guide repository organisation, digita l object management, technical infrastructure, and security, providing benchmarks for TRUSTworthiness, as exemplified by the INSDC.

Findability and indexing: Keywords and literature discoverability

Our comparative analysis of publication indexing systems further revealed patterns affecting article findability. MeSH terms offered broader coverage (69.9% vs. 47.2% for author keywords) and reduced redundancy that can increase literature discoverability. In contrast, author keywords frequently overlapped semantically, with terms such as “environmental DNA” and “eDNA” recorded separately, despite having identical meanings—a redundancy also observed in methodological descriptors. We recommend that researchers adopt hybrid search strategies combining MeSH terms and author keywords for systematic reviews while promoting standardised keyword vocabularies across the eDNA community. Consistent deposition of eDNA datasets with comprehensive metadata would mitigate many of these challenges, improving both data and literature accessibility.

Documentation and traceability of experimental protocols

Our LLM-based synthesis of 1,607 eDNA publications enabled large-scale identification of patterns in methodological reporting and data deposition. The analysis revealed several recurring features of current practice: most researchers place experimental data in Suppl. material 1 rather than the main text, and sample collection techniques vary widely across studies. DNA extraction generally follows established workflows, with commercial kits frequently highlighted as efficiency-enhancing alternatives. Data storage practices show a strong preference for GenBank (97.3% out of the 1,607 papers with repository information), with lower use of alternatives such as mBRAVE and FigShare. However, methodological transparency remains a major challenge, with only 0.4% of papers referencing protocols published in protocols.io. For instance, PCR conditions or at least PCR primer sequences should be consistently reported, as they are a critical determinant of taxonomic coverage, amplification bias, and comparability across eDNA studies. We also note that protocols.io is not the only platform through which protocols can be shared in a FAIR manner; alternative approaches may include the use of general-purpose repositories such as GitHub with DOI minting via Zenodo, Figshare, as well as other protocol-sharing platforms. Taken together, the large-scale synthesis revealed structural reporting weaknesses that extend beyond individual studies: methodological details hidden behind paywalls, inconsistent protocol documentation with minimal use of dedicated repositories, variable data accessibility practices, and systematic inconsistencies in research indexing. Together, poor transparency and inconsistent indexing limit the field’s ability to perform evidence synthesis and meta-analyses.

Consequences of poor metadata and incentive misalignment

Incomplete and inconsistent metadata reflect deeper structural imbalances between data producers and data users. While researchers generating data often prioritise publication timelines, downstream users depend on comprehensive, standardised metadata for reuse and synthesis. Limited training, inadequate resource allocation, and weak recognition of data stewardship exacerbate this imbalance. Although some journals have begun to implement formal guidance and metadata expectations to support FAIR eDNA data practices (e.g., the adoption of the FAIRe guidelines by Environmental DNA, as outlined in its Author Guidelines and Aims and Scope), enforcement and consistency across journals and disciplines remain variable. Consequently, incentives to invest substantial time in comprehensive metadata documentation are not always fully aligned with existing reward structures in research evaluation. However, the growing recognition of the importance and demand for data citation, as well as the need for AI-ready datasets (https://www.arxiv.org/pdf/2507.23018), suggests that improvement is possible, requiring aligned incentives that increase the value of metadata curation in research. Given the sensitivity of metabarcoding-derived taxa occurrences to operational flaws (e.g., contamination) and bioinformatic pipelines, the accessibility of metadata and the auditability of data processing workflows are particularly critical for DNA-based occurrence records in secondary repositories.

Towards coordinated, context-dependent standardisation

Transformative change in the eDNA landscape will require internationally standardised reporting frameworks that enhance methodological transparency, support reproducible bioinformatic analyses, improve interoperability, and ensure the long-term reusability of data and knowledge. These frameworks must address not only methodological reporting, but also consistent terminology, unified metadata structures, and indexing practices to enhance the discoverability of literature and data. Community-driven initiatives are beginning to emerge. For example, the FAIR eDNA (FAIRe) (Takahashi et al. 2025) project has developed a comprehensive metadata checklist that integrates existing standards with new terms tailored to eDNA workflows. These coordinated efforts demonstrate how standardisation and community involvement are essential to ensure that eDNA can achieve practical scientific and societal impact.

Conclusions

Our assessment of the European aquatic eDNA landscape underscores the urgent need for standardisation in methodologies and metadata reporting. While eDNA enables sensitive, non-invasive detection of invasive, rare, and indicator species, its full utility is currently constrained by inconsistent reporting, incomplete metadata, and fragmented data mobilisation across repositories. Key gaps exist in the documentation of DNA extraction methods, PCR protocols, sequencing platforms, and bioinformatic analysis, and significant regional disparities persist in the adoption of eDNA approaches, as well as infrastructure. Moving forward, coordinated efforts between field and laboratory method experts, data custodians, and policymakers are essential to embed minimum metadata

requirements into harmonised reporting frameworks and best-practice guidelines, ensuring that high-quality metadata becomes a natural outcome of methodological compliance. Importantly, such standardisation does not imply a single universal eDNA protocol; rather, it entails a portfolio of context-dependent, transparent, and well-documented methodological approaches that are comparable because they adhere to shared reporting principles. The adoption of FAIR, CARE, and TRUST principles will further ensure that eDNA data are accessible, interoperable, and ethically managed over the long term. Notably, the cost of poor metadata standardisation, in terms of wasted research investment, duplicated efforts, and reduced funding efficiency, was not quantified. Nevertheless, it represents a substantial hidden burden on the research ecosystem. By addressing these challenges, eDNA can become a cornerstone of aquatic and integrated biodiversity monitoring, enhancing conservation strategies, informing environmental policy, and enabling rigorous evaluation of restoration and management interventions.

Additional information

Conflict of interest

The authors have declared that no competing interests exist.

Ethical statement

No ethical statement was reported.

Use of AI

Generative AI tools were not used for the generation of the manuscript text. AI algorithms based on large language models (LLMs) and the associated computational pipeline were developed and used as part of the methodology.

Funding

This work was supported by the European Union under the Horizon Europe Programme (eDNAqua-Plan, Grant Agreement No. 101112800) and by the Norte Portugal Regional Operational Programme, under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). Furthermore, C.B. was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 857251, and E.E. was supported by the Research Council of Norway (contract number 342628/L10).

Author contributions

Conceptualization: PW, CB, EE, TL, FOC, MG, VN, KM, SP. Data curation: IK, EE, DK, FOC. Formal analysis: IK, PW, CB, EE, DK. Funding acquisition: MG, SP. Investigation: IK, PW, CB, EE, DK, MG, VN, SP. Methodology: IK, PW, CB, EE, DK, TL, FOC, MG, VN, SP. Project administration: NP, SP. Resources: KBS, NP, MG, SP. Software: IK, PW, CB, DK. Supervision: KBS, MG, SP. Validation: IK, PW, CB. Visualization: PW, CB, SP. Writing – original draft: IK, PW, CB, EE, TL, VN, KM, SP. Writing – review & editing: IK, PW, CB, EE, DK, TL, JP, KBS, NP, FOC, MG, VN, KM, SP.

Author ORCIDs

Ioannis Kavakiotis https://orcid.org/0000-0003-4240-3636
Peter Woollard https://orcid.org/0000-0002-7654-6902
Camila Babo https://orcid.org/0009-0004-7916-9833

Elianne Egge https://orcid.org/0000-0001-5790-6426
Dawid Krawczyk https://orcid.org/0000-0002-5276-7436
Tiina Laamanen https://orcid.org/0000-0003-2095-9916
Joana Paupério https://orcid.org/0000-0003-2569-0768
Karolina Bacela-Spychalska https://orcid.org/0000-0003-4498-5107
Nicolas Pade https://orcid.org/0000-0003-2733-9752
Filipe O. Costa https://orcid.org/0000-0001-5398-3942
Michal Grabowski https://orcid.org/0000-0002-4551-3454
Veera Norros https://orcid.org/0000-0002-7481-0693
Kristian Meissner https://orcid.org/0000-0001-6316-8554
Spiros Papakostas https://orcid.org/0000-0002-5563-0048

Data availability

The analysis code produced for the manuscript is available on GitHub (https://github.com/eDNAqua-Plan-project/european-aquatic-eDNA-landscape/) and archived via Zenodo (https://doi.org/10.5281/zenodo.18660070). All other data that support the findings of this paper are available in the main text and Suppl. material 1.

References

Abarenkov K, Nilsson RH, Larsson K-H, Taylor AFS, May TW, Frøslev TG, Pawlowska J, Lindahl B, Põldmaa K, Truong C, Vu D, Hosoya T, Niskanen T, Piirmann T, Ivanov F, Zirk A, Peterson M, Cheeke TE, Ishigami Y, Jansson AT, Jeppesen TS, Kristiansson E, Mikryukov V, Miller JT, Oono R, Ossandon FJ, Paupério J, Saar I, Schigel D, Suija A, Tedersoo L, Kõljalg U (2024) The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: Sequences, taxa and classifications reconsidered. Nucleic Acids Research 52: D791–D797. https://doi.org/10.1093/nar/gkad1039

Adams AJ, Kamoroff C, Daniele NR, Grasso RL, Halstead BJ, Kleeman PM, Mengelt C, Powelson K, Seaborn T, Goldberg CS (2024) From eDNA to decisions using a multi-method approach to restoration planning in streams. Scientific Reports 14: 14335. https://doi.org/10.1038/s41598-024-64612-5

Altermatt F, Couton M, Carraro L, Keck F, Lawson-Handley L, Leese F, Zhang X, Zhang Y, Blackman RC (2025) Utilizing aquatic environmental DNA to address global biodiversity targets. Nature Reviews Biodiversity 1: 332–346. https://doi.org/10.1038/s44358-025-00044-x

Beng KC, Corlett RT (2020) Applications of environmental DNA (eDNA) in ecology and conservation: Opportunities, challenges and prospects. Biodiversity and Conservation 29: 2089–2121. https://doi.org/10.1007/s10531-020-01980-0

Buttigieg PL, Morrison N, Smith B, Mungall CJ, Lewis SE, Envo Consortium (2013) The environment ontology: Contextualising biological and biomedical entities. Journal of Biomedical Semantics 4(1): 43. https://doi.org/10.1186/2041-1480-4-43

Cordier T, Alonso Sáez L, Apothéloz Perret Gentil L, Aylagas E, Bohan DA, Bouchez A, Chariton A, Creer S, Frühe L, Keck F, Keeley N (2021) [July] Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. Molecular Ecology 30(13): 2937–2958. https://doi.org/10.1111/mec.15472

Corrales X, Katsanevakis S, Coll M, Heymans JJ, Piroddi C, Ofir E, Gal G (2020) Advances and challenges in modelling the impacts of invasive alien species on aquatic

ecosystems. Biological Invasions 22: 907–934. https://doi.org/10.1007/s10530-019-02160-0

Dhammi IK, Kumar S (2014) Medical subject headings (MeSH) terms. Indian Journal of Orthopaedics 48: 443–444. https://doi.org/10.4103/0019-5413.139827

Duarte S, Simões L, Costa FO (2023) Current status and topical issues on the use of eDNA-based targeted detection of rare animal species. The Science of the Total Environment 904: 166675. https://doi.org/10.1016/j.scitotenv.2023.166675

Eloe-Fadrosh EA, Mungall CJ, Miller MA, Smith M, Patil SS, Kelliher JM, Johnson LYD, Rodriguez FE, Chain PSG, Hu B, Thornton MB, McCue LA, McHardy AC, Harris NL, Reddy TBK, Mukherjee S, Hunter CI, Walls R, Schriml LM (2024) A practical approach to using the Genomic Standards Consortium MIxS reporting standard for comparative genomics and metagenomics. In: In: Setubal JC, Stadler PF, Stoye J (Eds) Comparative Genomics. Methods in Molecular Biology, vol 2802. Humana, New York, NY, 587–609. https://doi.org/10.1007/978-1-0716-3838-5_20

Fonseca VG, Davison PI, Creach V, Stone D, Bass D, Tidbury HJ (2023) The Application of eDNA for Monitoring Aquatic Non-Indigenous Species: Practical and Policy Considerations. Diversity 15: 631. https://doi.org/10.3390/d15050631

Harper LR, Buxton AS, Rees HC, Bruce K, Brys R, Halfmaerten D, Read DS, Watson HV, Sayer CD, Jones EP, Priestley V, Mächler E, Múrria C, Garcés-Pastor S, Medupin C, Burgess K, Benson G, Boonham N, Griffiths RA, Lawson Handley L, Hänfling B (2019) Prospects and challenges of environmental DNA (eDNA) monitoring in freshwater ponds. Hydrobiologia 826: 25–41. https://doi.org/10.1007/s10750-018-3750-5

Hassenrück C, Poprick T, Helfer V, Molari M, Meyer R, Kostadinov I (2021) FAIR enough? A perspective on the status of nucleotide sequence data and metadata on public archives. bioRxiv 1–24. https://doi.org/10.1101/2021.09.23.461561

Hazen EL, Abrahms B, Brodie S, Carroll G, Jacox MG, Savoca MS, Scales KL, Sydeman WJ, Bograd SJ (2019) Marine top predators as climate and ecosystem sentinels. Frontiers in Ecology and the Environment 17: 565–574. https://doi.org/10.1002/fee.2125

Jerde CL, Chadderton WL, Mahon AR, Renshaw MA, Corush J, Budny ML, Mysorekar S, Lodge DM (2013) Detection of Asian carp DNA as part of a Great Lakes basin-wide surveillance program. Canadian Journal of Fisheries and Aquatic Sciences 70: 522–526. https://doi.org/10.1139/cjfas-2012-0478

Leese F, Altermatt F, Bouchez A, Ekrem T, Hering D, Meissner K, Mergen P, Pawlowski J, Piggott JJ, Rimet F, Steinke D, Taberlet P, Weigand AM, Abarenkov K, Beja P, Bervoets L, Björnsdóttir S, Boets P, Boggero A, Bones AM, Borja Á, Bruce K, Bursić V, Carlsson J, Čiampor F, Čiamporová-Zatovičová Z, Coissac E, Costa F, Costache M, Creer S, Csabai Z, Deiner K, DelValls Á, Drakare S, Duarte S, Eleršek T, Fazi S, Fišer C, Flot J-F, Fonseca V, Fontaneto D, Grabowski M, Graf W, Guðbrandsson J, Hellström M, Hershkovitz Y, Hollingsworth P, Japoshvili B, Jones JI, Kahlert M, Stroil BK, Kasapidis P, Kelly MG, Kelly-Quinn M, Keskin E, Kõljalg U, Ljubešić Z, Maček I, Mächler E, Mahon A, Marečková M, Mejdandzic M, Mircheva G, Montagna M, Moritz C, Mulk V, Naumoski A, Navodaru I, Padisák J, Pálsson S, Panksep K, Penev L, Petrusek A, Pfannkuchen MA, Primmer CR, Rinkevich B, Rotter A, Schmidt-Kloiber A, Segurado P, Speksnijder A, Stoev P, Strand M, Šulčius S, Sundberg P, Traugott M, Tsigenopoulos C, Turon X, Valentini A, van der Hoorn B, Várbíró G, Hadjilyra MIV, Viguri J, Vitonytė I, Vogler A, Vrålstad T, Wägele W, Wenne R, Winding A, Woodward G, Zegura B, Zimmermann J (2016) DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of

aquatic ecosystems in Europe. Research Ideas and Outcomes 2: e11321. https://doi.org/10.3897/rio.2.e11321

Leese F, Bouchez A, Abarenkov K, Altermatt F, Borja Á, Bruce K, Ekrem T, Čiampor F, Čiamporová-Zaťovičová Z, Costa FO, Duarte S, Elbrecht V, Fontaneto D, Franc A, Geiger MF, Hering D, Kahlert M, Kalamujić Stroil B, Kelly M, Keskin E, Liska I, Mergen P, Meissner K, Pawlowski J, Penev L, Reyjol Y, Rotter A, Steinke D, van der Wal B, Vitecek S, Zimmermann J, Weigand AM (2018) Why We Need Sustainable Networks Bridging Countries, Disciplines, Cultures and Generations for Aquatic Biomonitoring 2.0: A Perspective Derived From the DNAqua-Net COST Action. In: Bohan DA, Dumbrell AJ, Woodward G, Jackson M (Eds) Advances in Ecological Research. Next Generation Biomonitoring: Part 1. Academic Press, 63–99. https://doi.org/10.1016/bs.aecr.2018.01.001

Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, De Giusti M, L'Hours H, Hugo W, Jenkyns R, Khodiyar V, Martone ME, Mokrane M, Navale V, Petters J, Sierman B, Sokolova DV, Stockhause M, Westbrook J (2020) The TRUST Principles for digital repositories. Scientific Data 7: 144. https://doi.org/10.1038/s41597-020-0486-7

Meyer R, Appeltans W, Duncan WD, Dimitrova M, Gan Y-M, Jeppesen TS, Mungall C, Paul DL, Provoost P, Robertson T, Schriml L, Suominen S, Walls R, Sweetlove M, Ung V, de Putte AV, Wallis E, Wieczorek J, Buttigieg PL (2023) Aligning Standards Communities for Omics Biodiversity Data: Sustainable Darwin Core-MIxS Interoperability. Biodiversity Data Journal 11: e112420. https://doi.org/10.3897/BDJ.11.e112420

Nukazawa K, Hamasuna Y, Suzuki Y (2018) Simulating the Advection and Degradation of the Environmental DNA of Common Carp along a River. Environmental Science & Technology 52: 10562–10570. https://doi.org/10.1021/acs.est.8b02293

Prakash S (2021) Impact of climate change on aquatic ecosystem and its biodiversity: an overview. International Journal Biological Innovations 3(2): 312–317. https://doi.org/10.46505/IJBI.2021.3210

Priya AK, Muruganandam M, Rajamanickam S, Sivarethinamohan S, Gaddam MKR, Velusamy P, Gomathi R, Ravindiran G, Gurugubelli TR, Muniasamy SK (2023) Impact of climate change and anthropogenic activities on aquatic ecosystem - A review. Environmental Research 238: 117233. https://doi.org/10.1016/j.envres.2023.117233

R Core Team (2023) R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://doi.org/10.32614/r.manuals

Rourke ML, Fowler AM, Hughes JM, Broadhurst MK, DiBattista JD, Fielder S, Wilkes Walburn J, Furlan EM (2022) Environmental DNA (eDNA) as a tool for assessing fish biomass: A review of approaches and future considerations for resource surveys. Environmental DNA 4: 9–33. https://doi.org/10.1002/edn3.185

Sahu A, Kumar N, Pal Singh C, Singh M (2023) Environmental DNA (eDNA): Powerful technique for biodiversity conservation. Journal for Nature Conservation 71: 126325. https://doi.org/10.1016/j.jnc.2022.126325

Sayers E (2009) A General Introduction to the E-utilities. In: Entrez Programming Utilities Help [Internet]. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK25497/ [September 19 2025]

Simaika JP, Stribling J, Lento J, Bruder A, Poikane S, Moretti MS, Rivers-Moore N, Meissner K, Macadam CR (2024) Towards harmonized standards for freshwater biodiversity monitoring and biological assessment using benthic macroinvertebrates.

The Science of the Total Environment 918: 170360. https://doi.org/10.1016/j.scitotenv.2024.170360

Stewart KA (2019) Understanding the effects of biotic and abiotic factors on sources of aquatic environmental DNA. Biodiversity and Conservation 28: 983–1001. https://doi.org/10.1007/s10531-019-01709-8

Takahashi M, Frøslev TG, Paupério J, Thalinger B, Klymus K, Helbing CC, Villacorta-Rath C, Silliman K, Thompson LR, Jungbluth SP, Yong SY, Formel S, Jenkins G, Laporte M, Deagle B, Rajbhandari S, Jeppesen TS, Bissett A, Jerde C, Hahn EE, Schriml LM, Hunter C, Newman P, Woollard P, Harper LR, Dunn N, West K, Haderlé R, Wilkinson S, Acharya-Patel N, Lopez MLD, Cochrane G, Berry O (2025) A Metadata Checklist and Data Formatting Guidelines to Make eDNA FAIR (Findable, Accessible, Interoperable, and Reusable). Environmental DNA 7: e70100. https://doi.org/10.1002/edn3.70100

Takahashi M, Paupério J, Woollard P, Frøslev T, Thompson L, Silliman K, Formel S, Hunter CI, Babo C, Deiner K, Schriml L, Berry O (2026) A TDWG and GSC Collaborative Initiative To Develop an Environmental DNA (eDNA) Metadata Checklist. Biodiversity Information Science and Standards 10: e185867. https://doi.org/10.3897/biss.10.185867

Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, Bellemain E, Besnard A, Coissac E, Boyer F, Gaboriaud C, Jean P, Poulet N, Roset N, Copp GH, Geniez P, Pont D, Argillier C, Baudoin J-M, Peroux T, Crivelli AJ, Olivier A, Acqueberge M, Le Brun M, Møller PR, Willerslev E, Dejean T (2016) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Molecular Ecology 25: 929–942. https://doi.org/10.1111/mec.13428

Vaulot D, Sim CWH, Ong D, Teo B, Biwer C, Jamy M, Lopes dos Santos A (2022) metaPR^2: A database of eukaryotic 18S rRNA metabarcodes with an emphasis on protists. Molecular Ecology Resources 22: 3188–3201. https://doi.org/10.1111/1755-0998.13674

Yang H, Lee HJ (2018) Research Trend Visualization by MeSH Terms from PubMed. International Journal of Environmental Research and Public Health 15: 1113. https://doi.org/10.3390/ijerph15061113

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PSG, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, Gil IS, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber CL, Legg T, Ley RE, Lozupone CA, Ludwig W, Lyons D, Maguire E, Methé BA, Meyer F, Muegge B, Nakielny S, Nelson KE, Nemergut D, Neufeld JD, Newbold LK, Oliver AE, Pace NR, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman DA, Assunta-Sansone S, Schloss PD, Schriml L, Sinha R, Smith MI, Sodergren E, Spor A, Stombaugh J, Tiedje JM, Ward DV, Weinstock GM, Wendel D, White O, Whiteley A, Wilke A, Wortman JR, Yatsunenko T, Glöckner FO (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology 29: 415–420. https://doi.org/10.1038/nbt.1823

Supplementary material 1

Supplementary information

Authors: Ioannis Kavakiotis, Peter Woollard, Camila Babo, Elianne Egge, Dawid Krawczyk, Tiina Laamanen, Joana Paupério, Karolina Bacela-Spychalska, Nicolas Pade, Filipe O. Costa, Michal Grabowski, Veera Norros, Kristian Meissner, Spiros Papakostas

Data type: docx

Explanation note: Provides extended methodological and analytical details supporting the main text. Section A describes the large language model (LLM)-driven analysis of eDNA publications, including literature collection, model configuration, validation, and downstream data analysis. Section B includes a series of supplementary tables: (i) database property descriptions and evaluation framework; (ii) the stakeholder questionnaire on eDNA repositories; (iii) summaries of BioSample metadata coverage and search queries; and (iv) assessments of compliance with MIxS and ENVO standards. Together, these materials document the workflow, provide transparency in data collection and annotation, and highlight metadata reporting gaps across European aquatic eDNA resources.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.10.173612.suppl1